



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第三讲 线性回归模型

沈煜 博士 副教授

嘉定校区交通运输工程学院311室

yshen@tongji.edu.cn

2022年03月11日

计划进度



周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	爬虫
2	2022.03.04	沈煜	在线数据采集方法	
3	2022.03.11	沈煜	线性回归模型	
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	二元回归 (作业1)	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法 (作业2)	回归分析
8	2022.04.15	沈煜	关联: Apriori	
9	2022.04.22	沈煜	决策树、支持向量机 (作业3)	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络 (期末大作业)	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类: K-Means、DBSCAN	机器学习
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报 (1)	
17	2022.06.17	沈煜	期末汇报 (2)	

主要内容



- 引例
- 线性回归模型基础
- 参数估计
- 残差分析
- 统计检验
- 小结



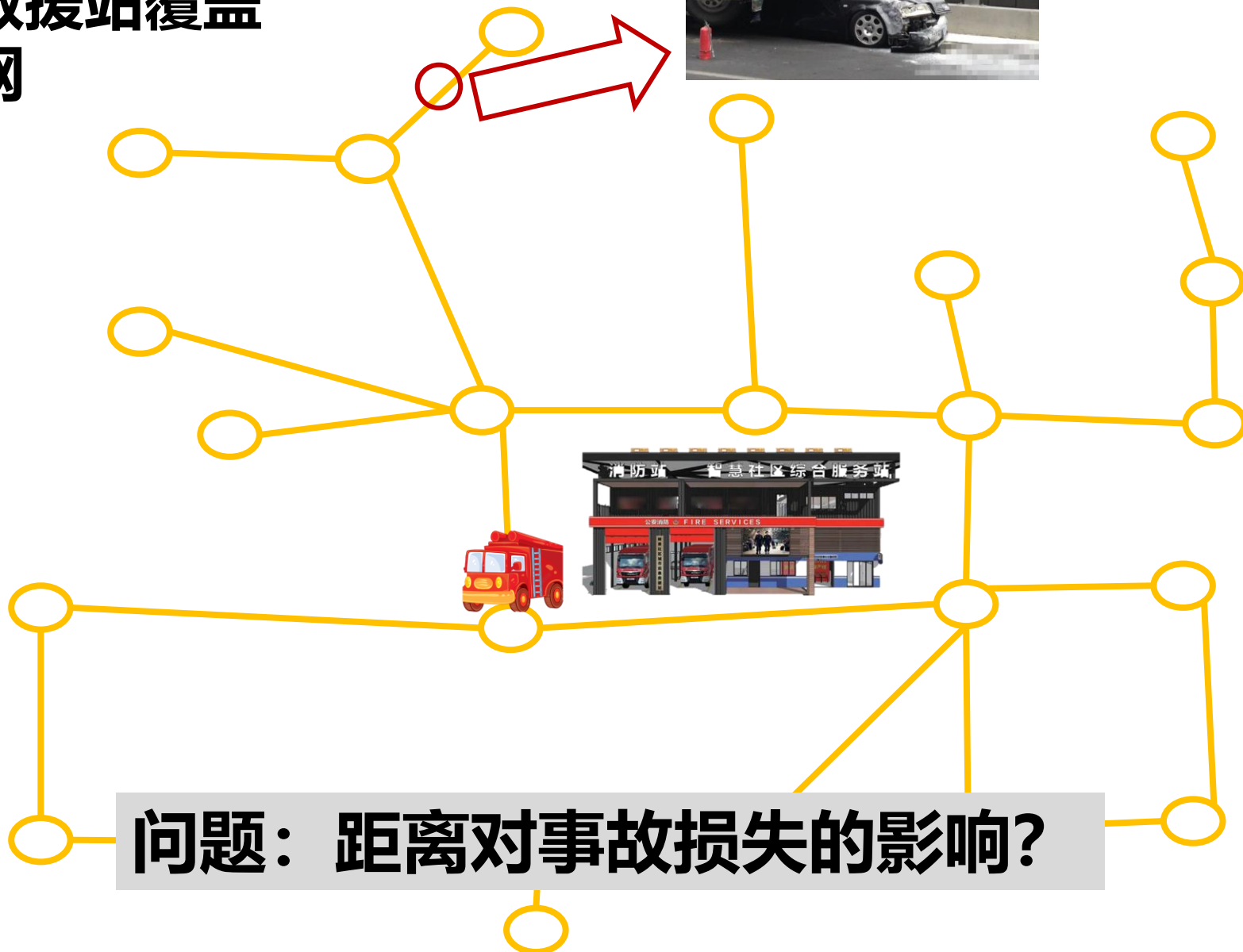
同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

引例

引例

应急救援站覆盖的路网

1. 医疗急救
2. 消防安全



问题：距离对事故损失的影响？

➤20起事故的损失和与最近应急站的距离

事故损失表

距离(km)	损失(千)*	距离(km)	损失(千)*
3.4	26.2	2.1	24.0
1.8	17.8	1.1	17.3
4.6	31.3	6.1	43.2
2.3	23.1	4.8	36.4
3.1	27.5	3.8	26.1
5.5	36.0	4.0	30.1
0.7	14.1	5.8	40.2
3.0	22.3	1.5	16.5
2.6	19.6	5.2	36.5
4.3	31.3	4.9	32.0

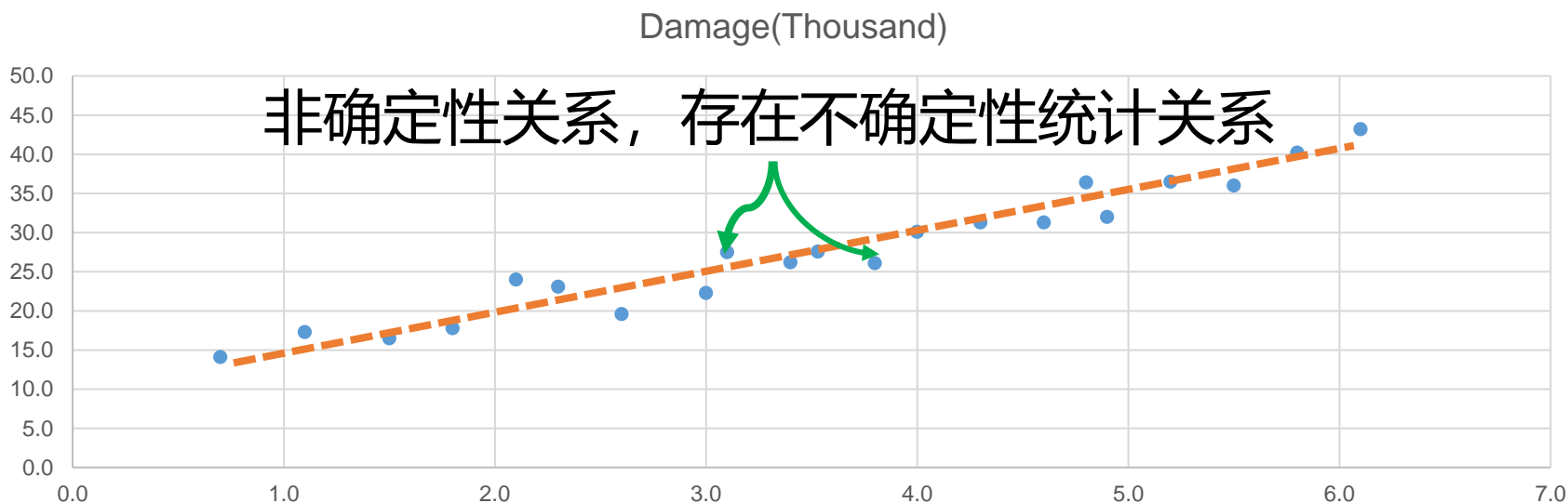
* 人和物损失

整体而言，距离越长，损失越大。
但也有问题：①间距并不是统一的；②有时候又出现相反的情况

引例



- 20起事故的损失和与最近应急站的距离
- 规律并不是那么明朗，放到一张散点图上



为了研究距离与损失之间的线性关系，考虑线性回归模型（一元）

解释：距离对损失的影响是多少？

预测：有一个新的距离，预测损失



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

线性规划模型基础

回归与回归方程



➤ 变量之间的相互关系

- **确定性关系**：当变量之间存在准确、严格关系时，用各种方程表示变量之间的函数关系，比如圆的周长和半径
- **不确定性的统计关系（相关关系）**：但是实际生活中，受其他因素干扰，许多变量之间并非严格的函数关系，不能用函数准确表示，这时称这种关系为回归关系
- 没有关系

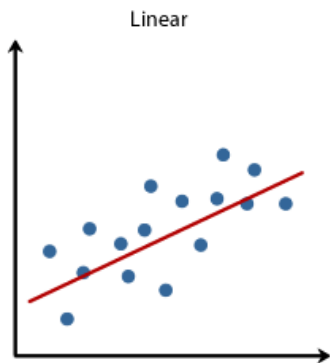
➤ **回归**：设法找出变量之间在数量上的依存变化关系

➤ **回归方程**：用函数表达式表达出来这个关系

回归与回归方程

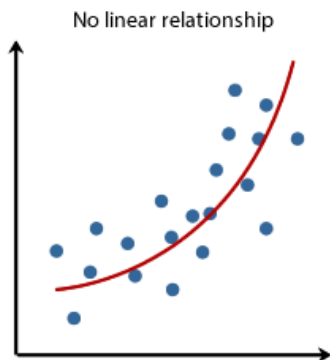


- 用直线方程式来表示这种关系叫做线性回归



$$y = \beta_0 + \beta_1 x + \varepsilon$$

- 用非直线方程式来表示这种关系叫做非线性回归



$$\log(y) = \beta_0 + \beta_1 x + \varepsilon \quad \text{对数-线性模型}$$

$$y = \beta_0 + \beta_1 \log(x) + \varepsilon \quad \text{线性-对数模型}$$

$$\log(p/(1-p)) = \beta_0 + \beta_1 x + \varepsilon \quad \text{比值-线性模型}$$

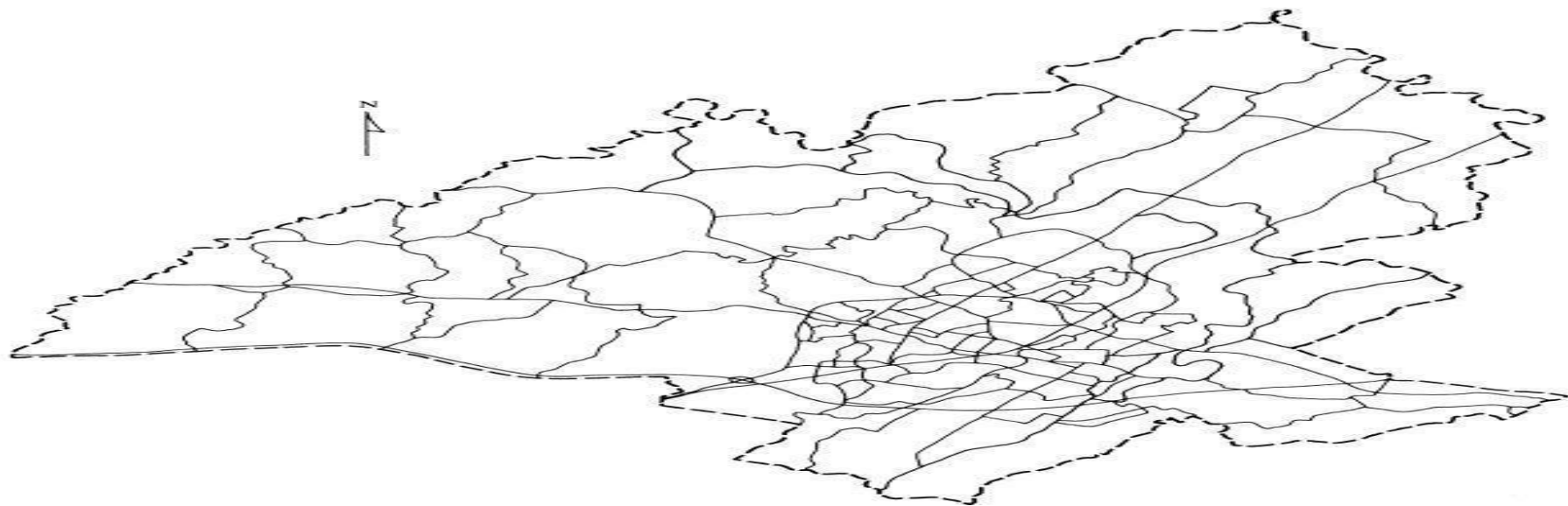
$$\log(x+1) \quad \log(y+1)$$

线性回归模型形式-总体回归模型



➤ 总体回归模型

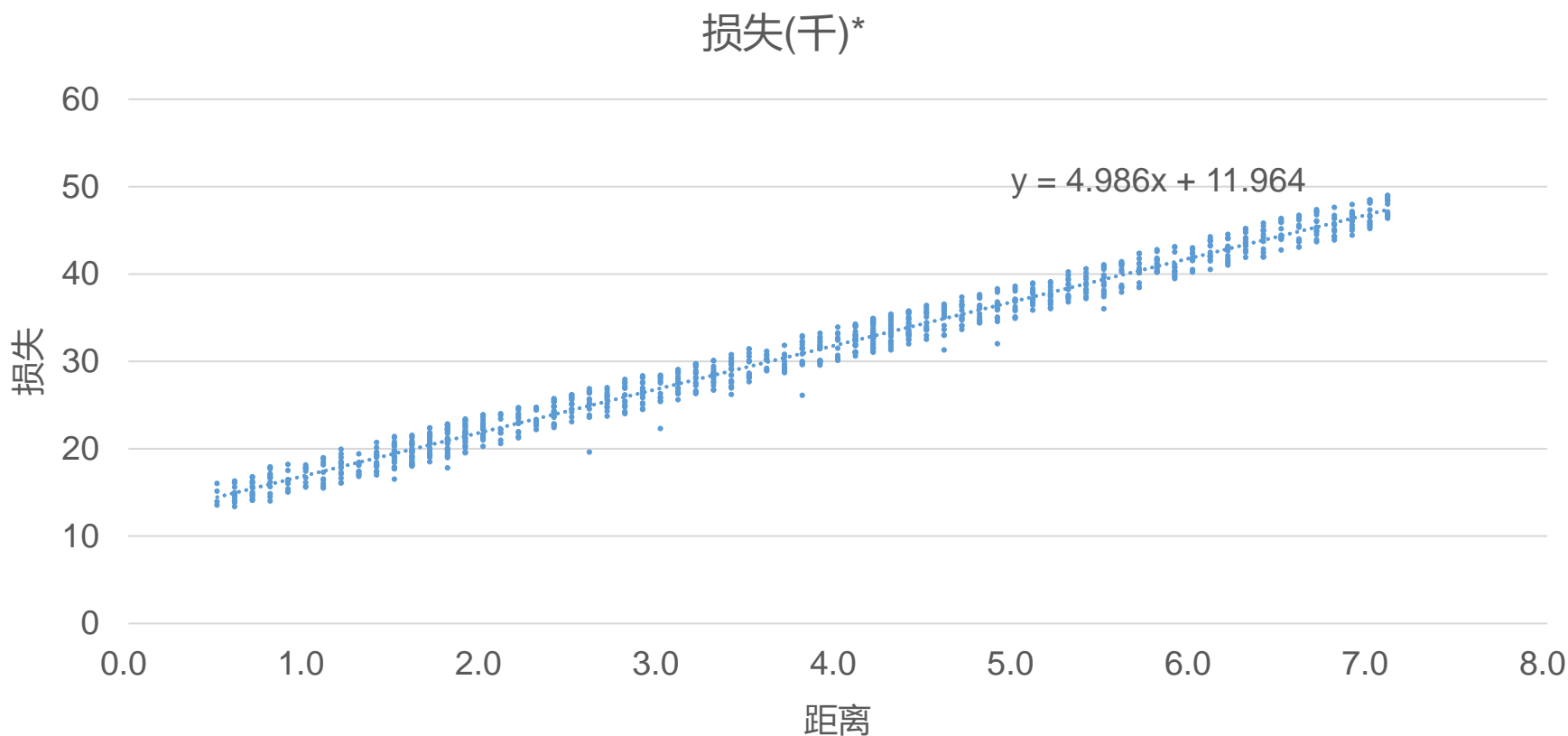
- 回归关心的是根据预测变量(x)的已知或给定值, 考察被响应变量的总体均值 $E(y|x) = f(x)$
- 线性回归模型的 $f(x)$ 采用线性函数的形式表示, 如 $f(x) = \beta_0 + \beta_1 x$
- 总体回归模型基于一个总体的数据构建模型, 如一个城市应急救援站覆盖范围内的事故情况



线性回归模型形式-总体回归模型



➤ 针对一个城市，一共105个救援小区，共1016起事故数据



线性回归模型形式-总体回归模型



➤ 随机误差项:

$$\varepsilon_i = y_i - E(y|x)$$

在给定x的情况下, y的条件期望值与其实值的区别。
即观测值与理论值的差

➤ 一元线性回归模型一般形式:

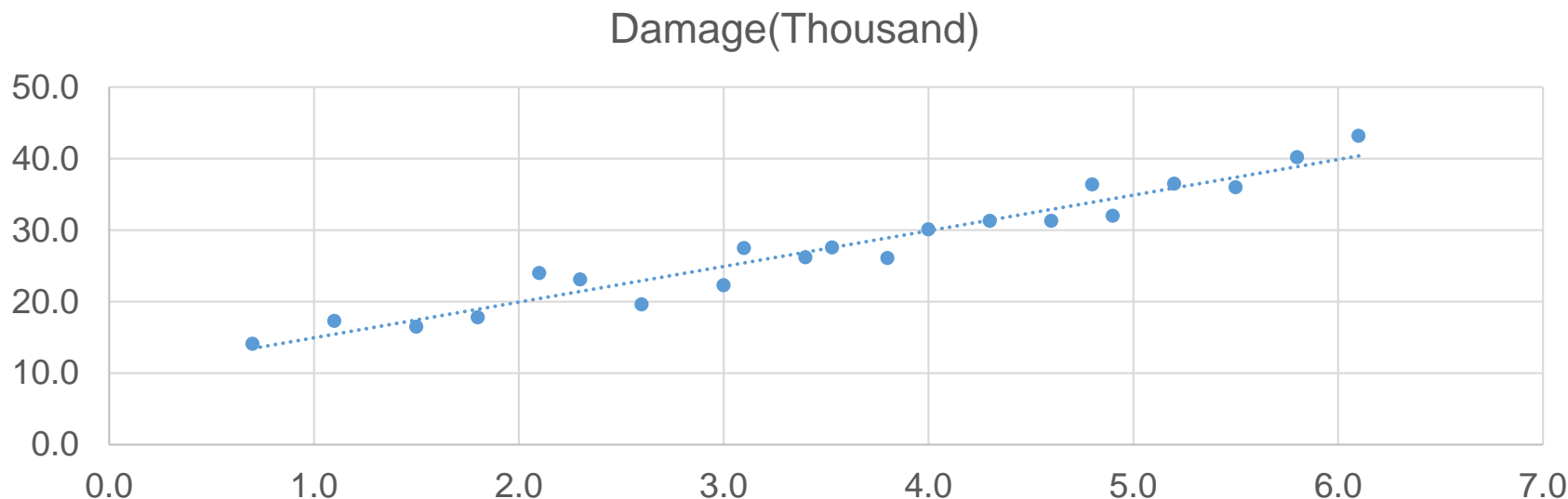
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i=1,2,\dots,n$$

y_i 等于条件期望加上随机误差项

线性回归模型形式-样本回归模型



- 在前面的总体中选择某个小区的数据，则其结果为



- 总体回归线基于的是全部的样本
- 但是，总体数据的收集总是耗费大量的人力物力
- 更多采用的是样本采样的形式，用样本数据估计回归函数
- 如：在前面的总体中从各组小区事故数据中各取一个进行观测

线性回归模型形式-样本回归模型

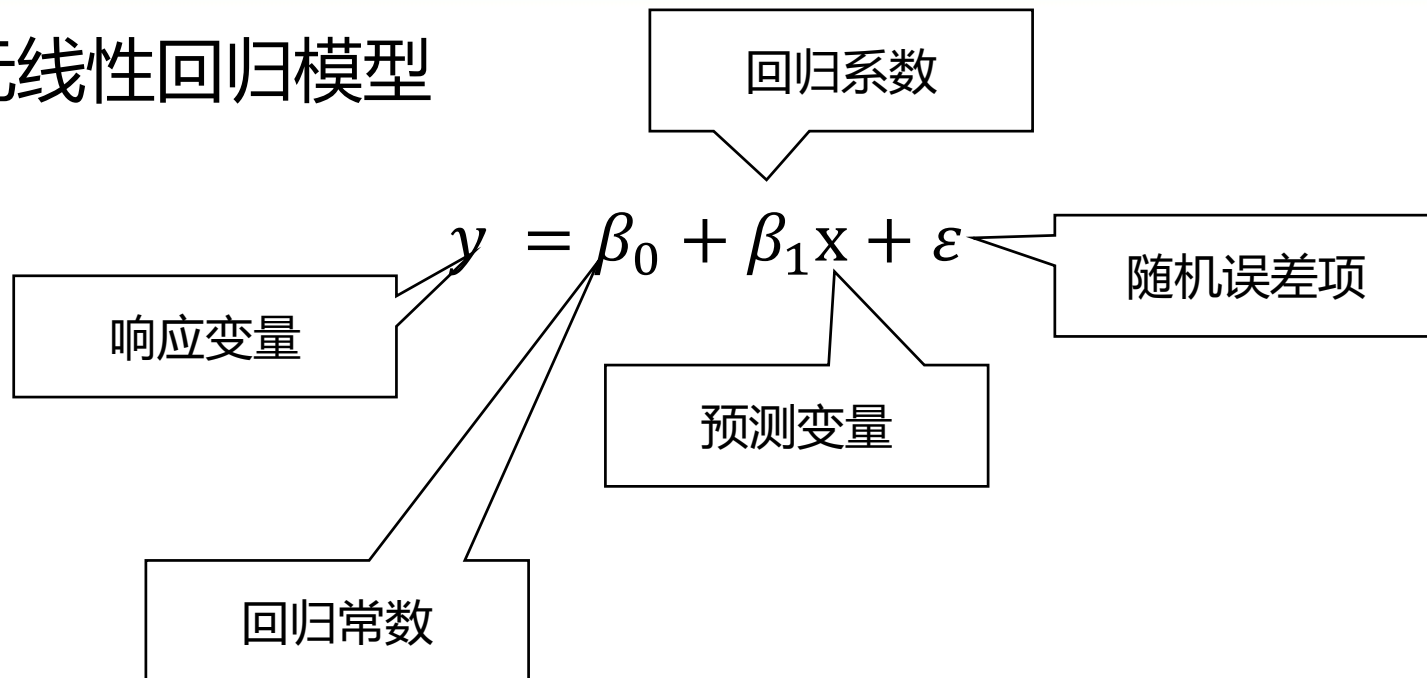


- 样本回归线：x与y散点图几乎成一条直线
- 样本回归函数： $E(y'_i | x'_i) = y'_i = \beta'_0 + \beta'_1 x'_i$
- 样本剩余项（残差）： $e_i = y_i - E(y'_i | x'_i) = y_i - y'_i$
- 则 $y_i = y'_i + e_i = \beta'_0 + \beta'_1 x'_i + e_i$
- 当样本的统计学规律与总体接近，那利用样本的回归模型能代替总体的回归模型

线性回归模型形式-一元



➤一元线性回归模型



给定样本观测 $\{(x_i, y_i): i = 1, 2, \dots, n\}$ 后, 上式可以写成

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

线性回归模型形式-多元



多元线性回归模型的一般形式

$$y_i = \beta_0 + \sum_{k=1}^K \beta_k x_{ki} + \varepsilon_i \quad i=1,2,\dots,n$$

- y_i 是第 i 个样本的因变量
- β_0 是截距, 为待估计的参数
- x_1, x_2, \dots, x_K 为自变量
- $\beta_1, \beta_2, \dots, \beta_K$ 为自变量的系数, 是待估计的参数
- ε_i 是第 i 个样本的随机误差
- 自变量系数 (β_k), 反映的是其他自变量不变情况下, 对应自变量每变化一个单位引起的因变量的变化

线性回归模型另一种形式



➤一元为例

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \beta_1 \bar{x} + \varepsilon_i$$

$$\beta'_0 = \beta_0 + \beta_1 \bar{x} = \bar{y}$$

最后, $y_i = \bar{y} + \beta_1 (x_i - \bar{x}) + \varepsilon_i$

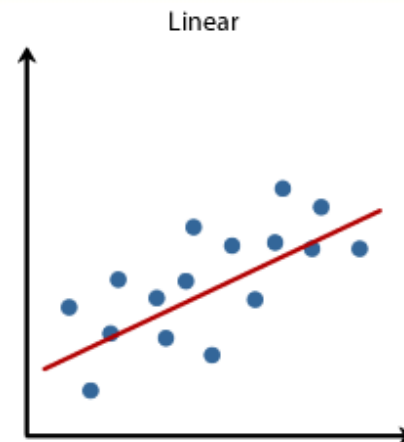
当没有任何模型或者 x 的时候, 我们会认为 y 的可能值是其均值

线性回归模型理论假设



➤对因变量的假设:

- 连续的
- 可以为负数
- 正态分布



➤对自变量的假设

- 有正确的期望函数，没有遗漏重要的自变量，没有多余的自变量（非高度相关、非无用）

➤因变量等于期望函数与随机干扰项之和

$$Y = X\beta + \varepsilon$$

线性回归模型理论假设



➤ 没有遗漏重要的自变量



地铁站附近当天自行车借用总量 (y)

$$R^2 = 0.24$$



当天的天气类型

周边土地利用强度，土地类型，公交车站距离，小区距离... ..

线性回归模型理论假设



➤ 预测变量之间不高度相关

➤ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

➤ 一个极端的例子, $x_1 = x_2$

➤ $y = \beta_0 + \beta_3 x_1 + \varepsilon$, 其中 $\beta_3 = \beta_1 + \beta_2$

➤ 对参数进行估计, 只需要 $\hat{\beta}_3 = \hat{\beta}_1 + \hat{\beta}_2$ 即可, 因为无法完全把 $\hat{\beta}_1, \hat{\beta}_2$ 分开, 无法根据现有数据去得到准确的估计值, 如果 $(\hat{\beta}_1, \hat{\beta}_2)$ 是一组估计, 那 $(\hat{\beta}_1 + 1, \hat{\beta}_2 - 1)$ 和 $(\hat{\beta}_1 + 2, \hat{\beta}_2 - 2)$ 等也是一组估计

➤ 即使预测变量和响应之间存在显著关系, 系数也可能看起来并不显著 (方差大)

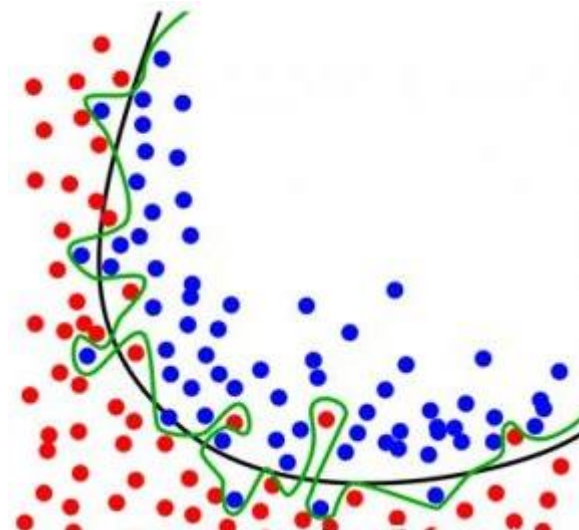
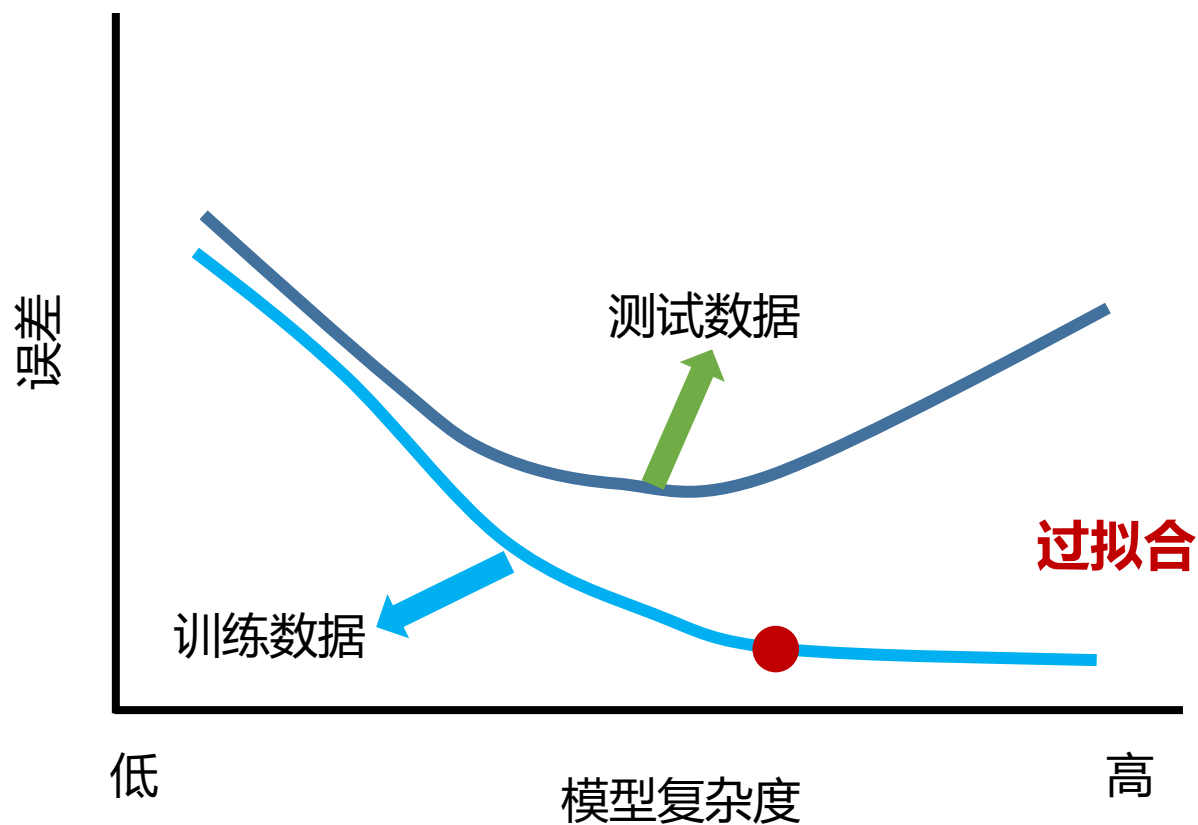
➤ 高度相关的预测变量的系数在样本之间差异很大。 (如 $\hat{\beta}_3 = 1$, $\hat{\beta}_1 = 10$, $\hat{\beta}_2 = -9$)

➤ 从模型中去除任何高度相关的项都将大幅影响其他高度相关项的估计系数。高度相关项的系数甚至会包含错误的符号。

线性回归模型理论假设



➤没有多余的自变量



线性回归模型理论假设



➤ 随机误差项假设（高斯-马尔科夫条件）

➤ 解释变量与随机误差项不相关

➤ 随机误差项服从正态分布，随机误差项具有0均值、同方差，且在不同样本点相互独立，不存在序列相关性

$$E(\varepsilon) = 0$$

$$\sigma^2(\varepsilon_i) = \sigma^2$$

$$\text{COV}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

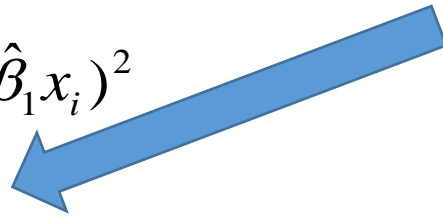
参数估计

参数 (β) 估计



- 普通最小二乘估计
- (Ordinary Least Square Estimation, 简记为OLSE)

最小二乘法就是寻找参数 β_0 、 β_1 的估计值使离差平方和**最小**

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$


$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 称为 y_i 的回归拟合值, 简称回归值或拟合值

$e_i = y_i - \hat{y}_i$ 称为 y_i 的残差

参数 (β) 估计



$$\begin{cases} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \left. \frac{\partial Q}{\partial \beta_1} \right|_{\beta_1 = \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^n e_i = 0 \\ \sum_{i=1}^n x_i e_i = 0 \end{cases}$$

经整理后,得方程组

$$\begin{cases} n\hat{\beta}_0 + \left(\sum_{i=1}^n x_i\right)\hat{\beta}_1 = \sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n x_i\right)\hat{\beta}_0 + \left(\sum_{i=1}^n x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^n x_i y_i \end{cases} \quad \rightarrow \quad \begin{cases} n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 = n\bar{y} \\ n\bar{x}\hat{\beta}_0 + \left(\sum_{i=1}^n x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^n x_i y_i \end{cases}$$

参数 (β) 估计



$$\begin{cases} \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \\ n\bar{x}(\bar{y} - \bar{x}\hat{\beta}_1) + \left(\sum_{i=1}^n x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^n x_i y_i \end{cases}$$



$$\begin{cases} \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)} \end{cases} \begin{matrix} \longrightarrow L_{xy} \\ \longrightarrow L_{xx} \end{matrix}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = L_{xy} / L_{xx} \end{cases}$$

$$L_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x} \bar{y}$$

$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2$$

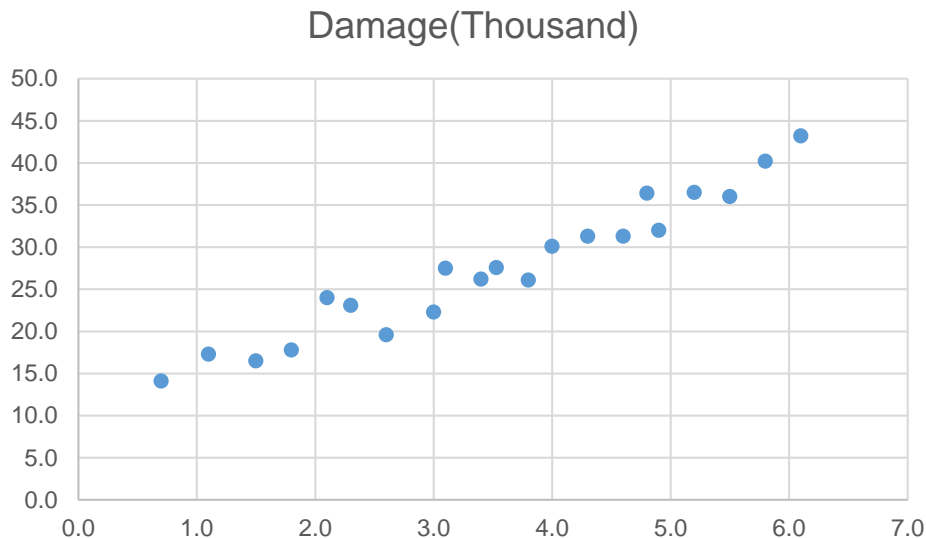
参数 (β) 估计——案例



➤ 20起事故的损失及事故发生地与最近的应急站的距离

事故损失表

距离(km)	损失(千)
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
0.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1
4.0	30.1
5.8	40.2
1.5	16.5
5.2	36.5
4.9	32.0



参数 (β) 估计——案例



$$\bar{x} = 3.53$$

$$\bar{y} = 27.58$$

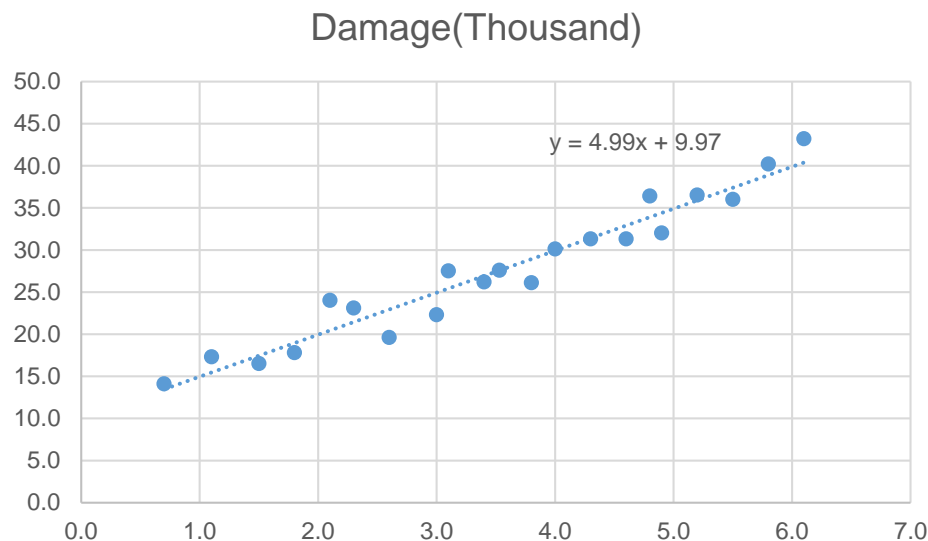
$$\begin{aligned} L_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ &= 299.10 - 20 \times 3.53^2 = 49.88 \end{aligned}$$

$$\begin{aligned} L_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \\ &= 2195.56 - 20 \times 3.53 \times 27.58 = 248.77 \end{aligned}$$

$$\beta_1 = L_{xy} / L_{xx} = 248.77 / 49.88 = 4.99$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 27.58 - 4.99 \times 3.53 = 9.97$$

$$y = 9.97 + 4.99x$$



距离对损失的影响是多少?

有一个新的距离(3km), 预测损失 24.94千

极大似然估计



➤ Maximum Likelihood Estimation

- 在给定样本的条件下，看成参数的函数，叫做**似然**
- 如果给定参数，把它看成样本的取值，那就是**概率**

$$\text{➤ } L(\hat{\beta}(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)) = \max L(\beta) = \max \prod_{i=1}^n f(\hat{y}_i | \hat{\beta})$$

➤ 其中 $L(\theta)$ 为似然函数， $\hat{\theta}(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ 为模型参数的极大似然估计（记为ML估计）

➤ 概率越大，事件发生的可能性就越大。就想找参数使得总体取样本的概率达到最大的参数，我们就叫做最大发生的可能性，那在这种情况下，参数 $\hat{\beta}$ 就是最优估计值

极大似然估计



在假设 $\varepsilon_i \sim N(0, \sigma^2)$ 时, 且独立的情况下, y_i 服从如下正态分布:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f_i(y_i)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2\right\}$$

$$\max \ln(L) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

min

y_1, y_2, \dots, y_n
的似然函数为:

对数似然
函数为:



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

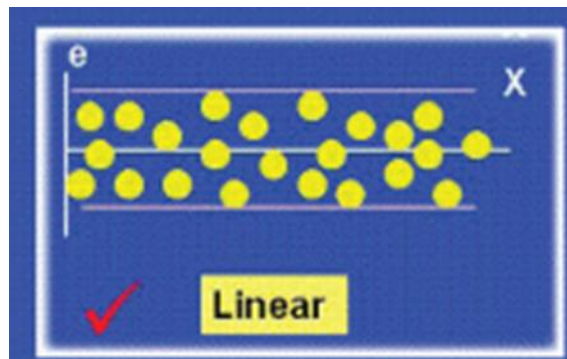
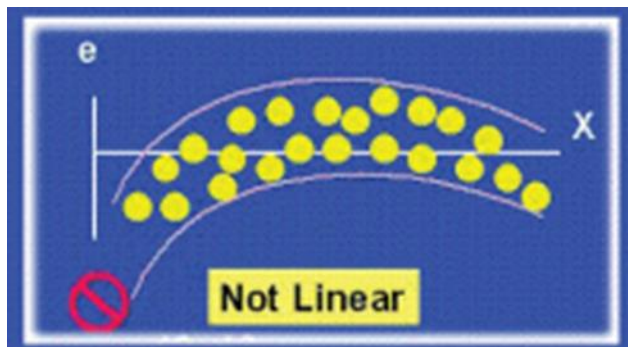
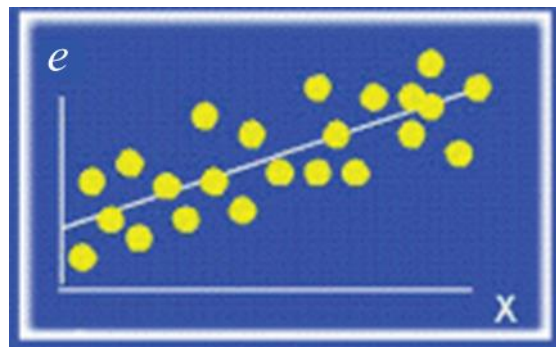
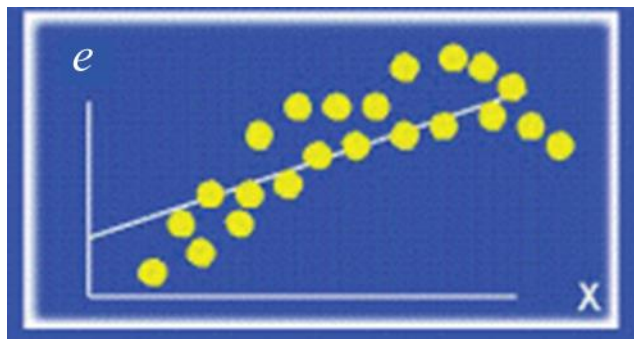
残差分析

- **残差图**：指以残差为纵坐标，以其他适宜的量为横坐标的散点图

- 最常见的横坐标选择是：
 - 1. 因变量的拟合值
 - 2. 某自变量的观察值
 - 3. 横坐标可取为观察时间或观察序号

线性

- 检验X和Y是否为线性关系，评估是否合乎线性回归成立的假设
- 绘图方法：使用残差的绘图分析（一个轴为残差 ε ，另一个轴为X或者Y），残差应该和另一个轴所代表的变量呈直线状



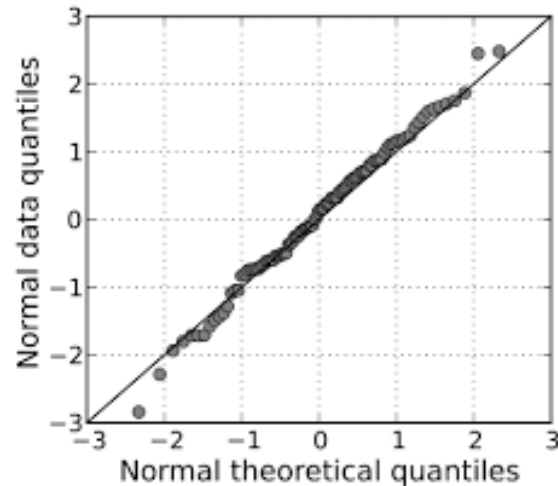
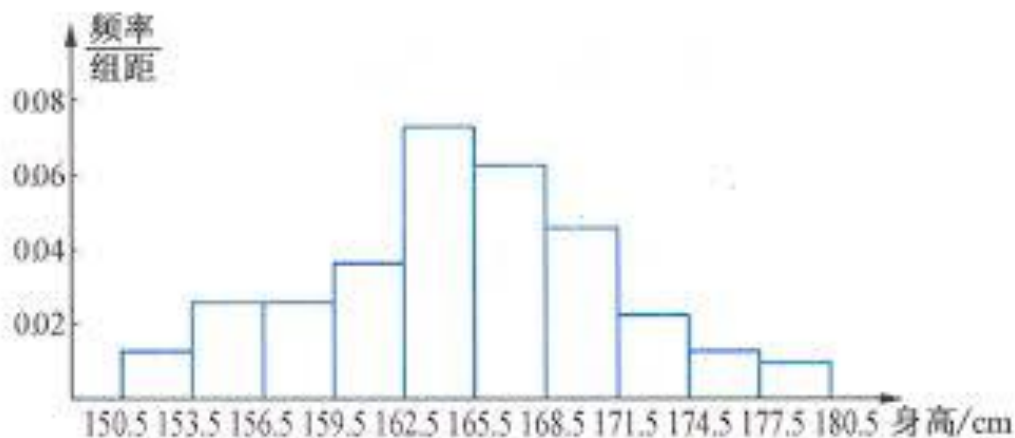
残差分析



➤ 正态分布

- 样本量足够，常态检验用算法比较好，但不够稳健
- 样本量不够，常态检验用算法的可信度不大

➤ 主要看图

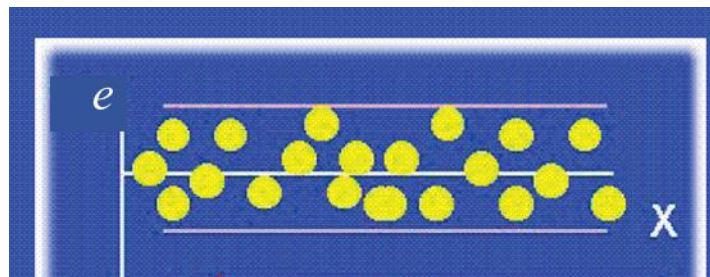
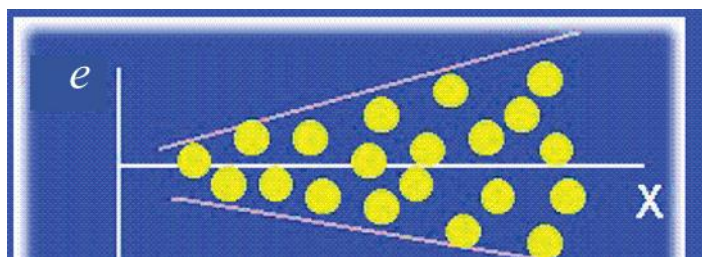
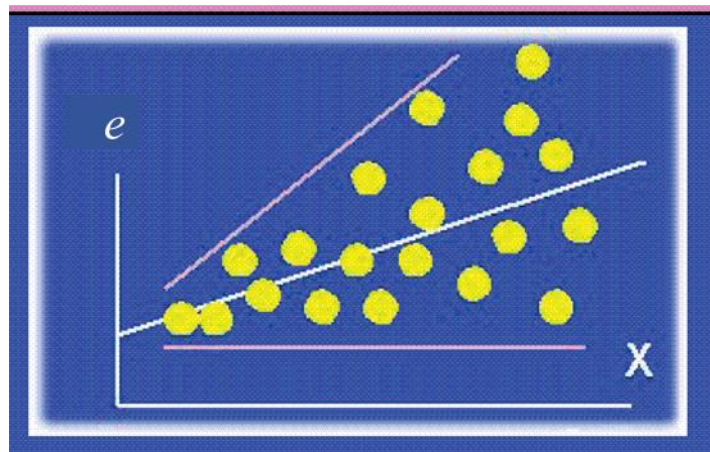
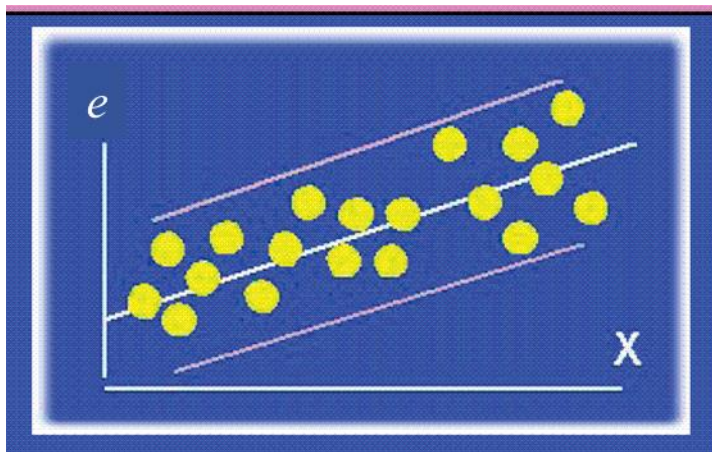


残差分析



➤ 方差齐性

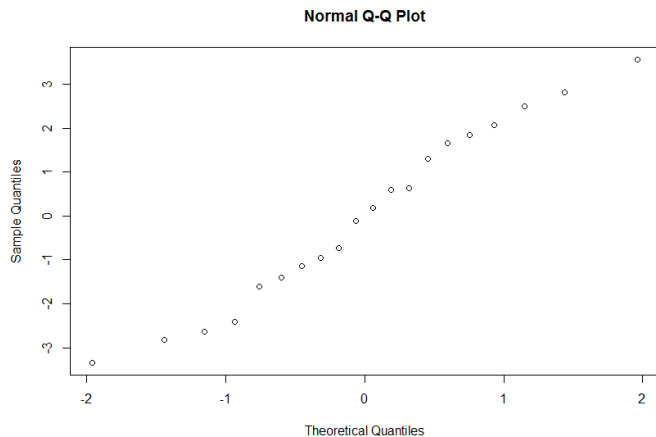
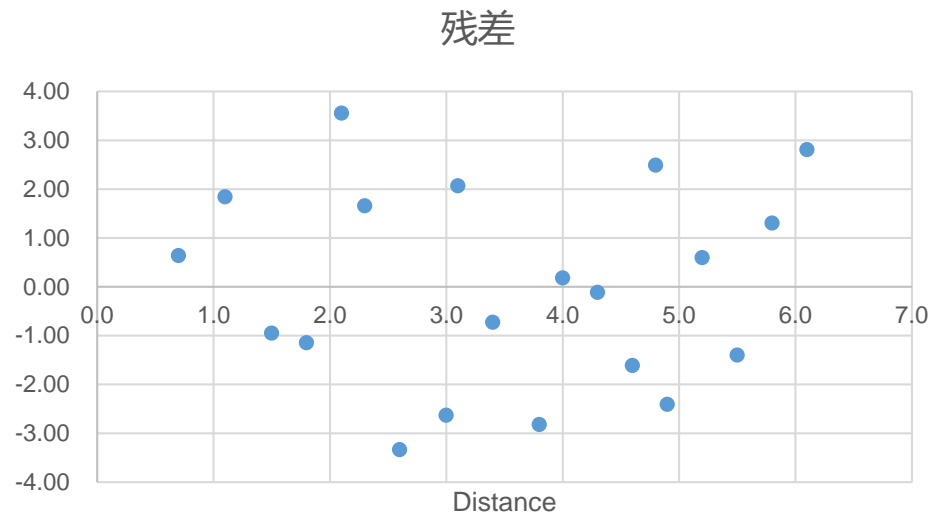
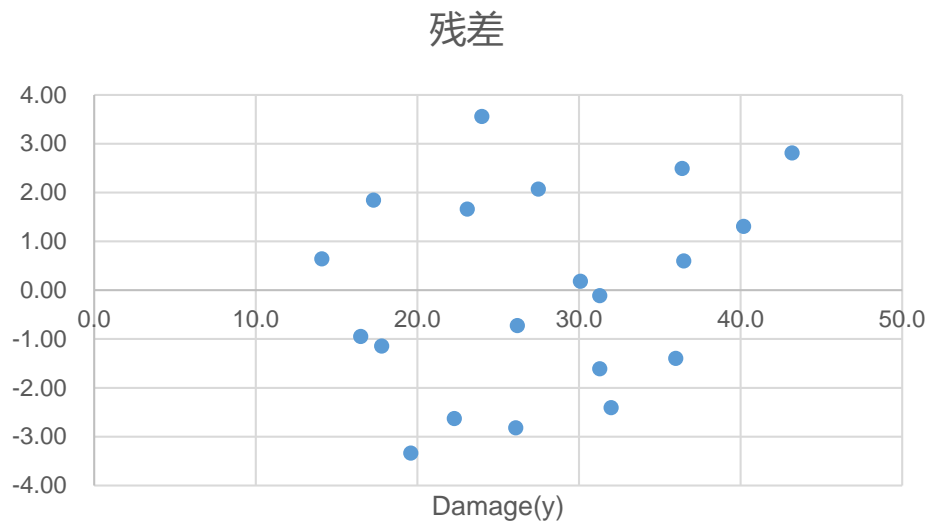
- 残差和X/Y轴来做竖轴和横轴
- 残差的方差一致性



残差分析——案例



➤ 事故损失模型残差



线性?
方差齐性?
常态分布?



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

统计检验

变量的显著性检验



- 回归分析是要判断解释变量 x 是否是被解释变量 y 的一个显著性的影响因素,即需要进行变量的显著性检验
- 变量的显著性检验所应用的方法是数理统计学中的假设检验
- 计量经计学中, 主要是针对变量的参数真值是否为零来进行显著性检验的

显著性检验



(1) 对总体参数提出假设

$$H_0: \beta_1=0, \quad H_1: \beta_1 \neq 0$$

(2) 以原假设 H_0 构造t统计量

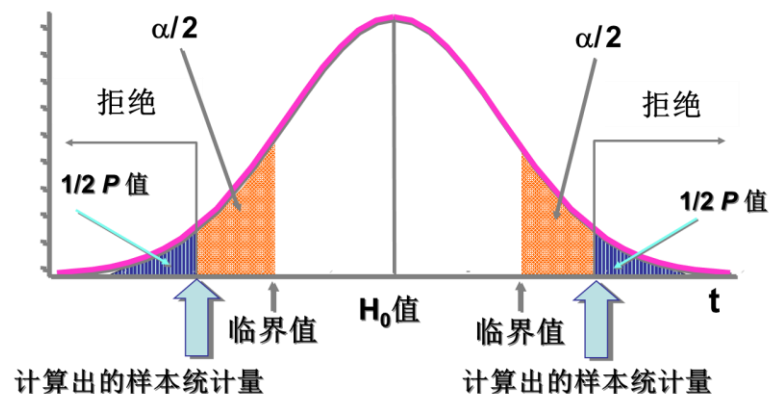
$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

(3) 给定显著性水平 α ，查t分布表，得临界值 $t_{\alpha/2}(n-2)$

(4) 比较，判断

若 $|t| > t_{\alpha/2}(n-2)$ ，则拒绝 H_0 ，接受 H_1 ；

若 $|t| \leq t_{\alpha/2}(n-2)$ ，则拒绝 H_1 ，接受 H_0 ；



拟合优度检验



拟合优度检验：对样本回归直线与样本观测值之间拟合程度的检验

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \\ \text{TSS} &= \text{ESS} + \text{RSS}\end{aligned}$$

TSS--总离差平方和

RSS--回归平方和

ESS—残差平方和

可决系数（判定系数） R^2 为：

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

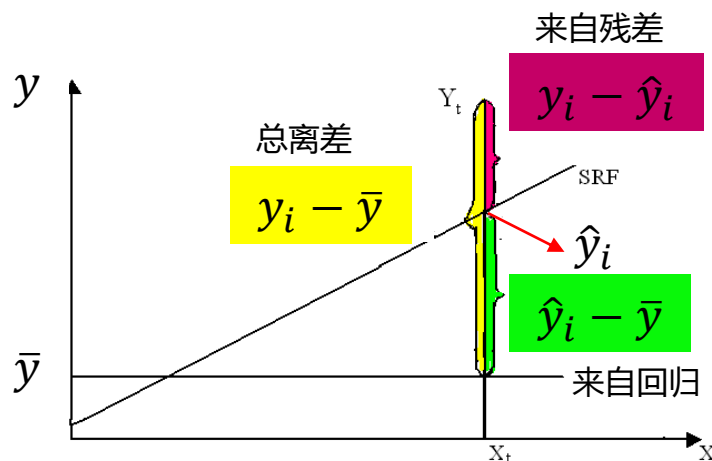
范围[0,1]

越靠近1，模型对数据的拟合程度越好

拟合优度检验



$$\triangleright y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = e_i + (\hat{y}_i - \bar{y})$$



如果 $y_i = \hat{y}_i$ 即实际观测值落在样本回归“线”上，则拟合最好。
可认为，“离差”全部来自回归线，而与“残差”无关。



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

小结

➤ 回归与回归方程

- 线性回归模型形式
- 模型理论假设

➤ 参数估计

- 最小二乘法
- 极大似然估计

➤ 残差分析

- 线性、正态、方差齐性

➤ 统计检验

- 变量的显著性检验
- 拟合优度检验



同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

第三讲 结束